Abstract for consideration for a presentation at the DH Benelux 2021 Conference

Title: Evaluating the multilingual capabilities of the OCCAM workflow: a case of digitised historical newspapers

Contributors:

Julie M. Birkholz, Sally Chambers, Michal Hradis, Pavel Smrz, on behalf of the EU funded OCCAM Project on OCR, Classification & Machine Translation.

Increased digitization of historical newspapers by cultural heritage institutions[1] has allowed humanities scholars to expand their corpora in terms of volume and diversity. This has accompanied an increase in the use of computational tools[2]. Simultaneously this has also brought to rise a new set of questions around the accuracy and value of this data and related studies. In this presentation we will focus on the multilingual challenge of studying historical newspapers from the lens of Belgium: questioning how a humanities researcher with an understanding of one language Dutch or French can accurately implement a study of a national press without the knowledge of the other most populous language. This question is core to the workflows being developed in the OCCAM (OCR, ClassificAtion & Machine Translation) project's digital humanities case. OCCAM implements a workflow for the integration of image classification, Translation Memories (TMs), Optical Character Recognition (OCR), and Machine Translation (MT) to support the automated translation of scanned documents.

We explain this through the case of the Belgian press using a set of multilingual historical newspapers from KBR- the Royal Library of Belgium's historical newspaper collections: BelgicaPress (Figure 1.). Through examples from a set of Dutch and French language newspapers from the early 1900s we explain how images of textual sources in multiple languages can efficiently be OCRed using the machine learning based model of PERO. In a subsequent workflow, the results of the OCR are then fed through a machine translation module. Originally developed for the machine translation of contemporary documents, we will report on the results using our digitized historical newspapers test case, to afford research of these historical documents.

---

[1] See for example the digitised newspapers in the Europeana: https://www.europeana.eu/en/collections/topic/18-newspapers

[2] Projects such as NewsEye (https://www.newseye.eu); Impresso (https://impresso-project.ch) and Living with Machines (https://livingwithmachines.ac.uk) explore computational approaches for analysis digitised historical newspaper corpora.

The OCCAM system results in high-quality OCR, that is flexible to diverse layouts of historical newspapers of varying quality (Figure 2.), with adaptation needed for line detection / layout identification (Figure 3.); and various combinations of printed and handwritten text (e.g. signatures in newspapers). The resulting text as a PAGE XML format affords an adaptable output for the translations, making this a flexible and adaptable tool for large multilingual collections that are in need of OCR as well as translations for researchers.

Figure 1. Multilingual example from within one newspaper article of De Standaard from 1919.

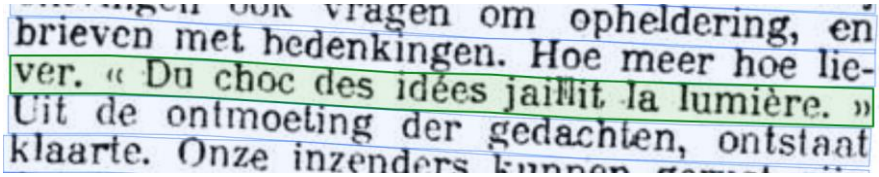

Figure 2. Signaling potential errors due to image quality issues

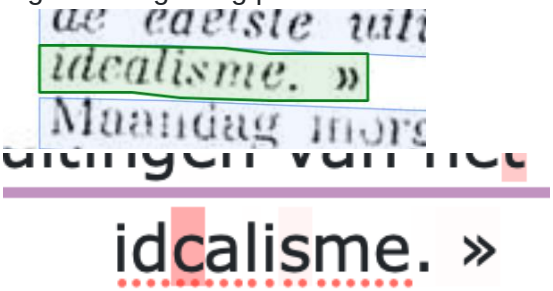

Figure 3. Layout identification